Statistical Considerations for Dominant Lethal Mutagenic Trials

by David S. Salsburg*

General Considerations

Statistics can be used in a biological assay as either a method of checking on the validity of conclusions or as a guide to the interpretation of experimental results. The division is sometimes made as one between hypothesis testing and estimation. However, the tool of hypothesis testing can be used as an interpretive aid in conjunction with the tools of estimation, so I prefer to make the divisions in terms of the uses made by the biologist.

If we are to apply statistical procedures to the dominant-lethal trail as a check of validity of conclusions, we would have to consider the inherent theoretical faults in the design. For instance, the animals actually treated constitute a small set of males and any conclusions must be conditional upon that set of males; thus, the probability levels computed are, themselves, random variables, functions of the random choice of males. Or, a typical trial will involve more than one test compound or dose of a test compound against the controls. Thus, any formal probability level computed must take into account questions of multiple comparisons. With comments like these, the statistician can sit in the marble halls of his floating island and throw bolts of doubt at the whole procedure. But, almost any biological assay can be made to suffer from this kind of criticism. At this stage in the development of mutagenicity testing, statistics can be much more effectively used to aid in the organization and interpretation of empirical results.

Three general problems face the experimenter in trying to organize and interpret the results of a dominant-lethal trial. He must be able to understand and display a vast amount of data (typically, 10 males per test group, 30 females per week, or 240 paired observations of implants and dead implants per group). He must have a running test procedure that will decide if a substance is to be suspected, a procedure with known probabilities of false positive and false negative results. He must be able to estimate the degree of an effect if he suspects one. A running test procedure will have to make use of the formal terminology of hypothesis testing. However, this is not a classical situation. For instance, in view of the inherent vagaries of the animals, it may be impossible to reduce the probability of false positives to a fixed constant (such as 5%) and the biologist may be forced to accept what he can, relying on statistical techniques to estimate the true running levels of error.

Review of the Literature

A detailed review of the literature from a statistical point of view is given elsewhere (*1*). Prior to 1970, statistics appear to have been used strictly as a check on validity. Chi-square contingency tables tend to be presented to compare counts of im-

*Pfizer Central Research, Medical Research Laboratory, Groton, Connecticut 06340.

plants in order to "prove" the obvious by computing significance levels of 1% or less. Attempts to organize the results to derive useful estimates of differential effect appear to be entirely without benefit of mathematical clergy.

In 1970, Krüger (*2*) attempted to interpret the results of dominant lethal trials in terms of a simple genetic mathematical model. He concluded that it was not possible to estimate the mutagenic effect, unconfounded by other factors, and he proposed no useful tools of analysis. He did not have available raw data from these tests and could only check the validity of his model againt mean counts in the published literature. On the other hand, Bishop (*3*) was able to use a large number of control animals from actual experiments. She was able to establish that the distribution of control data was sufficiently well behaved to allow for the use of standard robust statistical techniques. She set up a routine test method involving the use of a two-way analysis of variance with interaction [(weeks) × (treatment)] and introduced the use of variance stabilizing transforms to deal with counts of dead implants. She has kindly sent us a copy of her computer program. We have checked it against our data, and it appears to be a very well-written and well-documented program that can effectively handle the range of problems that might be expected in running a dominant-lethal trial.

The Bishop approach is much better than anything else appearing in the literature, but it is still far from ideal. In particular, it tests the overall mean levels of treatments and makes no provision for testing effects at specific stages of mutagenesis. In fact, a mild mutagen which affects the sperm during only one period will not produce a statistically significant treatment effect (although it might produce a significant interaction effect); and, thus, the method of testing chosen is ill-suited to the kind of alternative hypothesis one might expect in real life. Bishop also failed to consider what optimum combination of observations might work for a single mutagenic index. Instead, she chose to allow the analysis to run separately for counts of dead implants (estimating postimplantation losses) and for counts of total implants (estimating preimplantation losses).

Results of Investigations at Pfizer

At Pfizer, we were able to examine the data from over 4000 females taken during the control phases of more than 20 trials. From the analysis of this data we have developed an on-going method of computer analysis which has enabled us to estimate the true alpha level of our procedure, and we have been able to examine the value of our procedures with respect to known mutagens and compounds of unknown mutagenic potential. Details of our analysis of control data have been described elsewhere (*1*).

In general, we found that the number of implants for a given pregnant female mouse (all Charles River strain) can be effectively approximated by a binomial variate, as if there had been n implant sites, with each one having an independent opportunity to bear an implant with fixed probability p. Thus, the number of implants y_i found in the ith control female has a probability frequency of the form

$$\binom{n}{y} p^y (1-p)^{n-y}$$

The parameter p appears to be fixed at about 1/2, but the parameter n varies from one lot of females to another, ranging between 22 and 26.

If the number of implants can be fitted to a binomial $(n;p)$, then the total number of implants in M pregnant females will also be a binomial with parameters $(Mn;p)$. Furthermore, if we let the occurrence of dead implants be a set of independent Bernoulli variables, conditional on the occurrence of an implant, with probability of death, r, then X_i, the number of dead implants in the ith female, has a conditional frequency of the form

$$\binom{Y_i}{x} r^x (1-r)^{Y_i-x}$$

From these theoretical considerations, it can be shown that the total number of dead implants in M pregnant females has an unconditional binomial distribution with parameters $(Mn;pr)$ or a frequency of the form

$$\binom{Mn}{x} (pr)^x (1-pr)^{Mn-x}$$

If we assume that the effect of a mutagen will be to change the parameters p (the probability of an implant) or r (the conditional probability of a death, given an implant), then the number of implants and dead implants will continue to have a binomial distribution, even with a treatment effect. Thus, the arcsine transformation will stabilize the variance for both treatment and controls, regardless of the effect of treatment. This is not true of the square-root transformation chosen by Bishop, since that variance stabilization will hold only as long as the probability of a dead implant (r) remains small. By the arcsine transform, we mean

$$Z = 2 \text{ arcsine} \left\{ \begin{matrix} \text{number of implants}/nM, \text{ or} \\ \text{number of dead implants}/nM \end{matrix} \right\}^{1/2}$$

Regardless of the underlying probabilities, this transform has a variance $1/nM$. In our running method of analysis, we chose $n = 24$. It is clear that the variances will remain stabilized even if we have misestimated the value of n, as long as the ratio under the radical sign remains less than 1.0.

We have also examined the patterns of change that occurred over time among our control animals. There is a clear indication that the control parameters change with time. Figure 1 illustrates these changes. The upper part of the figure displays the average number of implants per pregnant control female (an estimate of $n/2$ in our binomial model if we assume p is constant at 0.5) across the entire 8 weeks of specific trials. The lower part of the figure displays the arcsine transform of the mean numbers of dead implants. There is an apparent cyclic pattern in the number of implants, with peaks occurring in September and valleys in April. There is also a significant ($p<0.01$) upward trend in the number of dead implants.

Methods of Analysis Now Being Used at Pfizer

We now have a running computer program which is written in Fortran but is mildly bound to the specific input/output configurations of the PDP–10 computer. Copies of the set of programs are available to anyone who requests them, with the understanding that some minor changes will have to be made in the flow of data. The program produces four pages of output for each treatment group. The first page lists the daily counts of pregnant females, numbers of implants, and numbers of dead implants that form the basic input, along with appropriate ratios and 3-day subtotals. This enables the experimenter to see gross and obvious patterns at a glance and to check for transcription errors in the initial input data.

The second page displays mean levels of implants, numbers of pregnant females, implants, dead implants, living implants, and ratios of these for each of the 8 weeks of trial. This enables the investigator to see the entire eight weeks of a single treatment group together, to gain subjective or "gut feeling" insights.

The third page is of the kind displayed in Figure 2. This is a plot of mean daily levels of a given measure (one of the z statistics described in the previous section or one of the more sophisticated second-order moment indices described in the next section) against a regression plotted from the mean daily control values. This regression, based on controls is an important part of the running analysis. In order to increase the power of the test, the entire 8 weeks of control values are compared against a single week of treatment values. Early experience with the trial indicated that the mean levels for controls tended to change over the 8-week period, so it was inappropriate to compare the treatment values for a single week against the overall mean of the controls. Instead, we fit a linear regres-

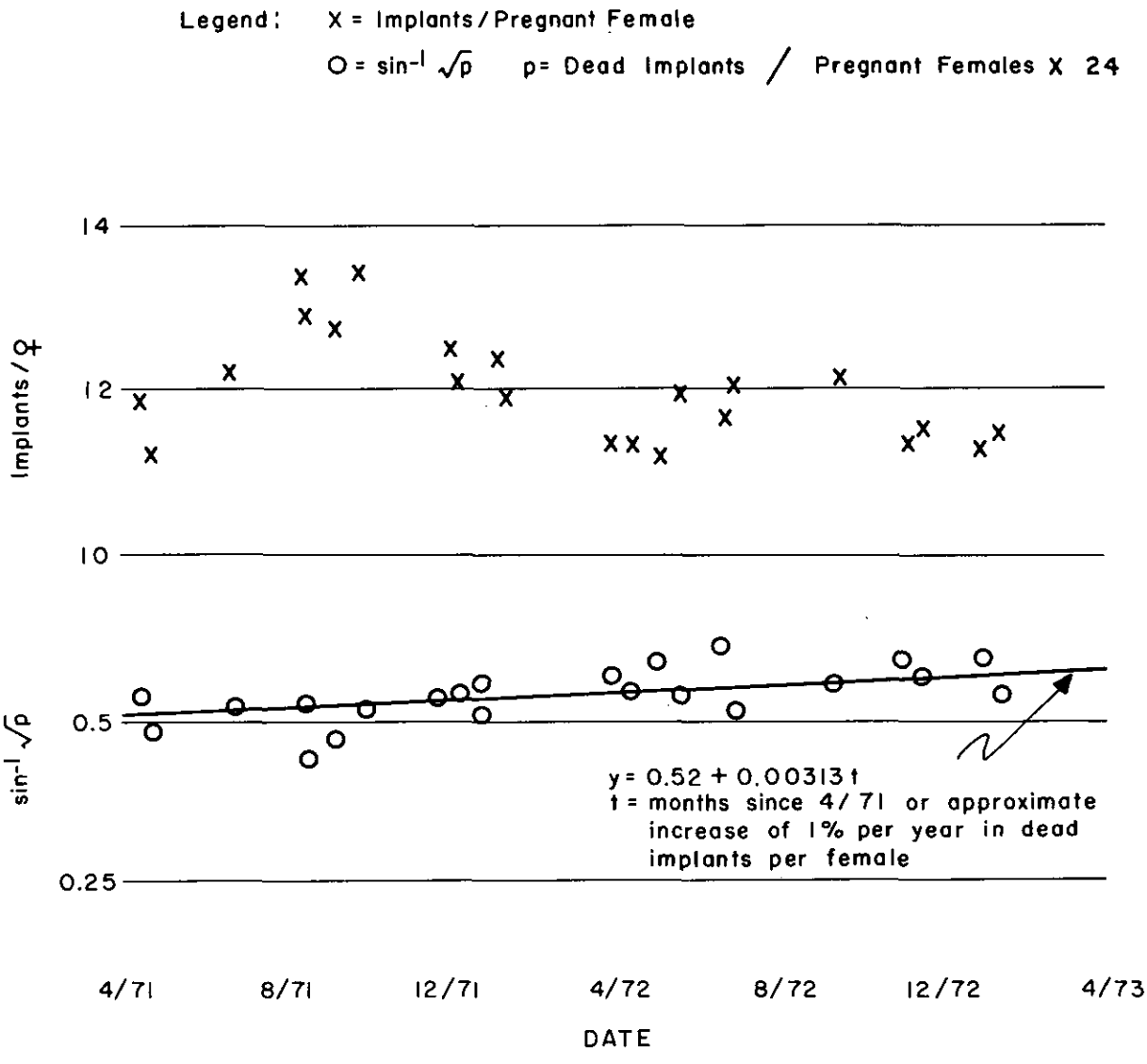

FIGURE 1. Control vs. experience over 2 years: (×) implants/pregnant female; (○) $\sin^{-1}\sqrt{p}$, where p = dead implants/ pregnant females × 24.

sion against day of trial to the controls and test a given week of treatment against that regression. The computer plot displays 90% and 95% tolerance bands above the control regression along with that estimated regression. Thus, the investigator can see graphically how and to what extent the treatment values were beyond those predicted from control for a given week.

The final page of output displays the results of statistical tests comparing the mean treatment levels for each week against the estimated control regression. In these tests, we ignore the variance of the controls, and we run a test which is conditional on the estimated regression. Furthermore, we run

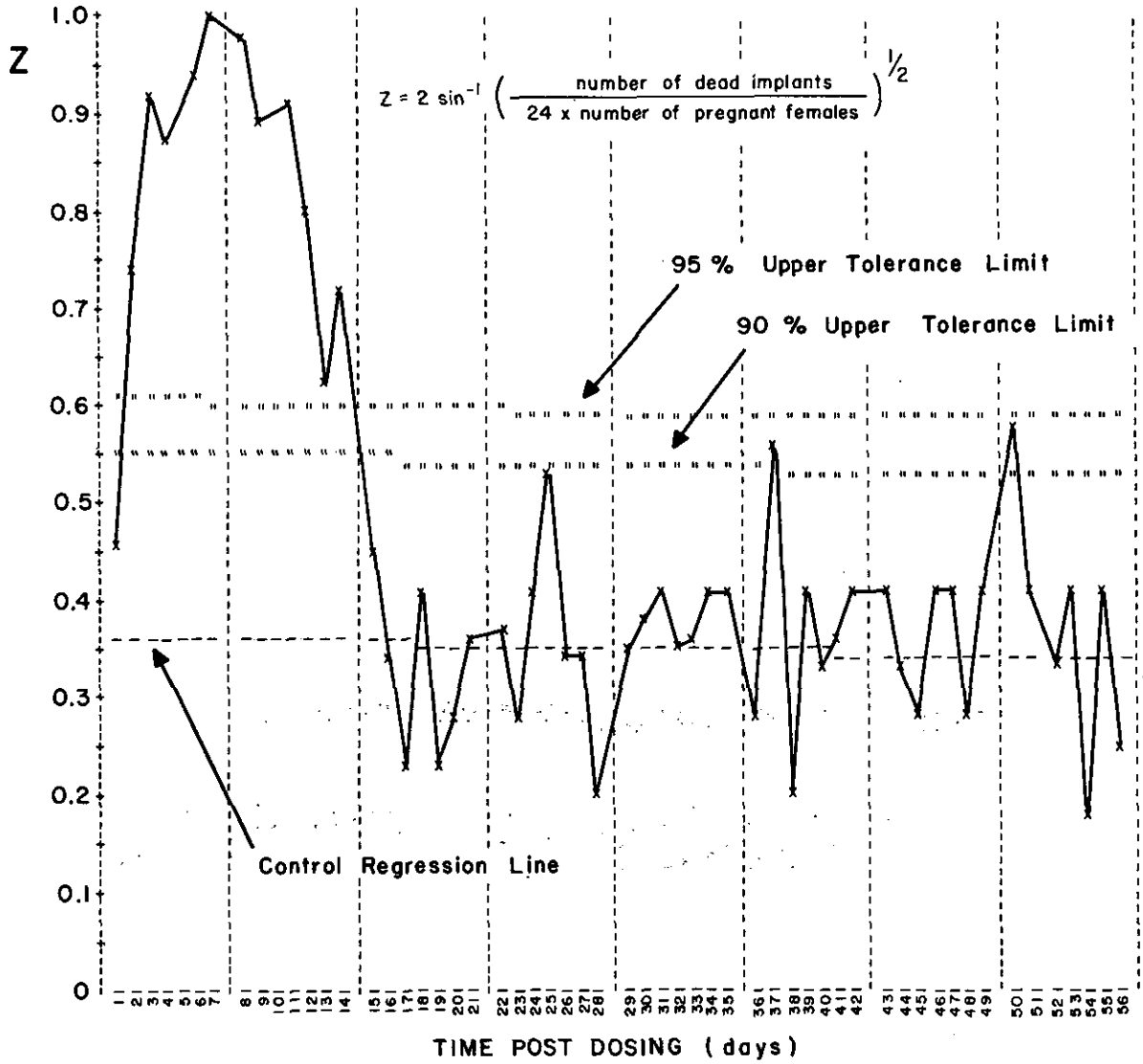

FIGURE 2. Control regression line and transformed EMS–360 data.

eight separate tests, one for each week. We use a nominal 1% level to test each week. The true alpha level is, of course, greater, being a function of the control variance and multiplicity of tests. Were all the tests independent and the control data without variance, the overall significance level would be approximately 8%. Our experience indicates that we are running at about 9% false positives.

For theoretical purposes, we have computed upper bounds on the control variance and attempted to estimate the power of this test (*1*). Estimates of the power are plotted in Figure 3. The vertical axis represents the power or probability of detecting a change

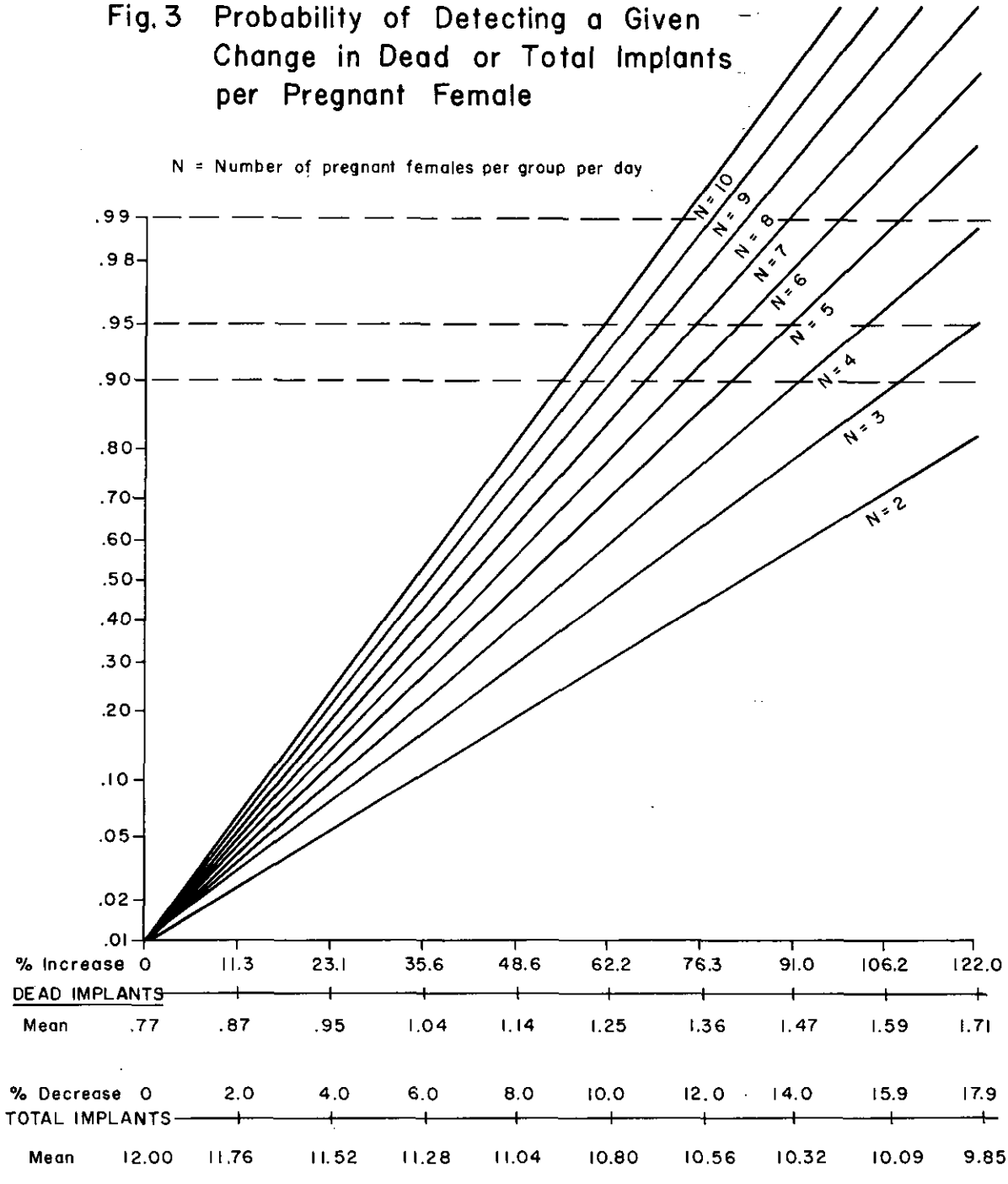

FIGURE 3. Probability of detecting a given change in dead or total implants per pregnant female.

of the magnitude indicated on the horizontal axis. Each ray represents a fixed number of females per group impregnated each day. This figure suggests, for instance, that we can be 99% sure of detecting a doubling in the mean number of dead implants per female if we run at six pregnant females a day.

Constructing a Single Mutagenic Index

The actual daily observations from the dominant lethal mutagenic trial consist of a two-dimensional vector,

$$\begin{pmatrix}\text{number of implants}\\ \text{number of dead implants}\end{pmatrix}$$

It should be possible to derive from these two numbers a single mutagenic index that will cover both pre- and postimplantation losses. In fact, a great deal of the pre-1970 literature deals with just this question. If we fall back upon the binomial model proposed above it can be shown that, if X_i = number of dead implants for the ith female and Y_i = total number of implants for the ith female, then

$$\begin{aligned} E(X) &= npr \\ E(Y) &= np \\ Var(X) &= npr(1-pr) \\ Var(Y) &= np(1-p) \\ Cov(X,Y) &= npr\ (1-p) \end{aligned}$$

The mean number of dead implants per pregnant female estimates $E(X)$, and the mean number of implants per pregnant female estimates $E(Y)$. It does not seem possible to find ratios or simple linear combinations of these two estimates that will be an increasing function of both pre- and postimplantation losses. Table 1 displays various indices based upon these two estimates, with the appropriate combinations of parameters of which they are consistent estimators. The columns labeled "conditions" show that for some mutagenic conditions they will tend to remain constant or actually decrease.

It would appear that any attempt to construct a consistent estimator of a combination of parameters that will increase for both pre- and postimplantation losses will have to involve second order moments like the variances or covariance. With clever enough juggling of the formulae for expectation, variance, and covariance, a number of such indices can be found. Two of these indices are displayed in Table 1. If we use the sample moments of the data, we can construct moment estimators of such indices. These are consistent estimators, but they may be biased, and it might be possible to find more efficient ones by means of maximum likelihood computations.

However, the present state of the art is

Table 1. Indices of mutagenic activity.[a]

Index		Parameter estimated	Conditions			
Verbal	Symbolic		$r\rightarrow$ $p\downarrow$	$p\rightarrow$ $r\uparrow$	$r\uparrow$ but $pr=k$	$r\rightarrow$ $p\uparrow$
Mean number of dead implants	$\overline{x}$	npr	$\downarrow$	$\uparrow$	$\rightarrow$	$\uparrow$
Number of dead implants/number of living implants	$\overline{x}/(\overline{y}-\overline{x})$	$pr/(1-pr)$	$\downarrow$	$\uparrow$	$\rightarrow$	$\uparrow$
Number of dead implants/number of implants	$\overline{x}/\overline{y}$	r	$\rightarrow$	$\uparrow$	$\uparrow$	$\rightarrow$
Number of dead implants—number of living implants	$\overline{x}-(\overline{y}-\overline{x})$	$n(2pr-1)$	$\downarrow$	$\uparrow$	$\rightarrow$	$\uparrow$
(prob of death/number of implants)	$S_{xy}/(S_y^2\,\overline{y})$	r/np	$\uparrow$	$\uparrow$	$\uparrow$	$\downarrow$
(prob of death) × (unconditional prob of living)	$S^2_x/\overline{y}$ * [b]	$r(1-pr)$	$\uparrow$	$\uparrow$	$\uparrow$	$\downarrow$

[a] Definitions:

p = Prob implant; r = Prob death, given an implant; n = number of implant sites; X_i = number of dead implants, ith female; Y_i = number of implants, ith female; m = number of pregnant females; $r\uparrow$ => implantation loss; $p\downarrow$=> implantation loss.

$\overline{x} = \Sigma X_i/m$

$y = \Sigma Y_i/m$

$S_x^2 = \Sigma(X_i-\overline{x})^2/(m-1)$

$S_y^2 = \Sigma(Y_i-\overline{x})^2/(m-1)$

$S_{xy} = \Sigma(X_i-\overline{x})\ (Y_i-\overline{y})/(m-1)$

$E(X) = npr \qquad \mathrm{Var}(X) = npr(1-pr)$

$E(Y) = np \qquad \mathrm{Var}(Y) = np(1-p)$

$\mathrm{Cov}(X,Y) = npr(1-p)$

[b] $L_n\ (S_x^2/\overline{y})$

such that we should first find a useful index that can make sense to the biologist. So, in our first tentative attempts to locate such an index, we have restricted attention to these moment estimators. It would appear, from our first few runs, that the estimator, ln $(S_y^2/\bar{y})$ is the best of those tried, best in the sense that it will declare statistical significance for known mutagens and fails to call significance for many of the situations we have identified as false positives using mean number of dead implants (or its arcsine transform).

Acknowledgements

A great deal of the work behind this paper is due to the efforts of Dr. Verne Ray and Leon Just of Pfizer Central Research who have joined with me in writing a more definitive paper (*1*). In addition to contributing a great deal of the statistical and mathematical back-up, Mr. Just is responsible for the running computer program described here. Dr. Ray has provided the impetus, the biological insights, and the general air of sensibility behind the work reported here. Furthermore, Miss Martha Hyneck, who has overseen the development of and actually controlled the running or our dominant-lethal mutagenic trials over a 2½ year period deserves the credit for many of initial insights that lead to our analyses of data and for the amazing amount of careful hard work, without which the numbers we analyzed would never have been available.

REFERENCES

1. Just, L. J., Ray, V. A., and Salsburg, D. S. Statistical analysis of the dominant lethal mutagenic assay in the mouse, submitted for publication.
2. Krüger, J. Statistical methods in mutation research. In: *Chemical Mutagenesis in Mammals and Man.* F. Vogel and G. Rohrborn, Eds. Springer-Verlag, New York, 1970.
3. Epstein, S. S., et al. Mutagenic and antifertility effects of TEPA and METEPA in Mice. *Toxicol. Appl. Pharmacol.* 17: 23. (1970).